**CS6140 Final Report**

**ABSTRACT**

The machine learning field can be briefly described as enabling machines to make sound decisions using previous experiences. This has exhibited an impressive development in recent years due to the increased processing power of computers. Therefore, it is essential to further understand and learn the basics of machine learning. This paper report discusses the findings and motivations from the final project for CS6140 Machine Learning. The dissatisfied customers, advancement of air travel and increase in people mobility have led to the development of sophisticated machine learning algorithms to improve the experience. The report details the motivation for investigating airport satisfaction data, analysis on dataset features and an investigation into model analysis via the various tools introduced to us over the course of the summer 2022 semester. We first provide the introduction of this report and provide a brief background of the motivation. Then the data and implementation are discussed to review the fundamental concepts for machine learning using SVM, Naïve Bayes and Logistic Regression. Then we point out the results of designing machine learning experiments and their performance evaluation. Finally, we provide the group's conclusion for this final report.

Keywords: Machine Learning, Support Vector Machines, Feature Analysis, Naïve Bayes, Logistic Regression, Kernel

**INTRODUCTION**

As the world slowly emerges from the Covid-19 pandemic, travel is reentering the lives of millions of Americans. How can airline companies capitalize on this return to travel and make the experience as positive as possible so industry can quickly recover? We are going to use machine learning, particularly the support vector machines (SVM) method, to explore what makes an airline trip positive and satisfying. This paper explores what key features an airline should invest in and use for prediction and qualify their relationship between each other. It is essential to understand the needs and comfort level of a multitude of customers. Therefore, customer feedback is the key for any airline industry to prosper and advance. There are multitude of possible ways of collecting customer response and one traditional way is getting feedback during the actual journey. But this method sometimes is not appealing to customer by filling in feedback forms. Another form of feedback is through using the online website or mobile application during booking of a trip or after completion of a journey, an email with the link for feedback can be sent to the passenger. One better approach is to use social media like Facebook and Twitter to request feedback. The customers are asked to rate (1 for poor and 5 for excellent) the experience and service. Because of the enormous amount of data available, machine learning made it possible to analyze these huge datasets to highly predict and classify the models. In this report, machine learning techniques such as SVM are used to analyze the class satisfaction sentiment classification model for the survey data provided.

**DATA**

The public “Airline Passenger Satisfaction” dataset from Kaggle was utilized for this investigation (1). The data set contained 25 possible features and over 1500 data entries for a robust training and test set (1). Data was encoded as integer values. Missing values were placed in median values or class means. The link for the dataset can be downloaded from this [website](https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction). Each column on the dataset will be the identified features for the customer feedback and the label would be the column “satisfaction”. The label will be either satisfied or dissatisfied(neutral) for the customer response. Most of the features were rated from 1 as poor to 5 for excellent. Other features that were continuous namely age, flight distance, departure delay and arrival delay are measured. Using this data can determine factors that are highly correlated to a satisfied customer.

**IMPLEMENTATION**

To effectively track the progress of our SVM models, we decided to initially train two baseline models: a Naïve Bayes and a logistic regression model. We chose Naïve Bayes and Logistic Regression to get our initial baseline results because they're fast and easy to use. These models are the “go to examples” for generative and discriminative classifiers, respectively. Naïve bayes uses Bayes theorem to make predictions based on training set data (3). By using this model, we are establishing a baseline with simple probabilistic calculations.

In order to investigate various decision boundaries, scikit-learn was utilized to implement a SVM with linear, polynomial and radial-basis-function kernels (2). SVMs can be described as a “hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks”. By transforming the feature-space through kernels, using the Kernel-Trick, SVMs can be applied to various decision boundary shapes. Various visualizations were presented to illustrate the message and overall results.

**RESULTS**

A Naïve Bayes implementation which treated 0-5 ratings as categorical features resulted in the following prediction statistics: Accuracy 0.886, Precision 0.8968692449355433, Recall 0.8935779816513761 and F1 score 0.8952205882352942. We used these metrics as a baseline performance for analysis.

The Logistic Regression model was able to take the data prepared with the integer encodings to perform training on the same features. A 10-fold cross validation was performed to reveal an accuracy of 0.8735. Logistic Regression assumes that “the variable follows a binomial distribution of the linear combination of the input variables” which may not be true of all features used during training (4). This is one possible explanation for a slightly worse performance than Naïve Bayes. An alternative explanation may be due to overfitting the model.

With baseline results from Naïve Bayes and Logistic Regression combined, we can establish a baseline expectation of at least 0.88 accuracy.

We began our analysis utilizing SVMs with Linear Kernels. This linear boundary was explored through 2, 3 and all features. The linear boundary, as shown by comparing Figure 1 to Figure 2, was extremely similar across feature combinations for 2 feature experimentation. This, linked with similarly low accuracy results ranging from 68-80% which were significantly below the baseline, led us to conclude high correlation between features (Table 1). High correlation between features suggests a potentially dependent relationship between features, breaking the underlying premise of many of the models we intended to investigate. Similar conclusions were made when investigating the 3-feature combination space as accuracy results were not significantly improved upon by increasing the features space.

When using all features in linear kernels, the best trained model resulted in an accuracy of 0.872, which approached our baseline but did not improve upon it. An interesting note from our data revealed that decreasing the slack variable C for values 0.1, 1, 10 and 100 did not assist in improving the performance. Accuracy remained approximately 0.873 across all these experiments. It is likely that the solution could not be encapsulated by a linear decision boundary and the SVM constraints were not the limiting factor.

Our suspicion of limiting a linear boundary placed on the model was confirmed by our investigation into polynomial and RBF Kernels. Polynomial kernels performed with an average of 0.938 and RBF kernels performed with an average of 0.952.

Despite the success of the RBF Kernel model, our suspicions of dependent data led us to calculate a correlation heat map (Figure 2). This heat map certainly revealed correlation possibilities which made intuitive sense as humans reviewing the map. For example, late departure and late arrival times were highly correlated. Both features described “lateness”. However, most features were observed to be relatively uncorrelated. Fears developed of potentially dependent data were put at ease when it was observed that reducing the feature space results in worse performance by the RBF and Polynomial kernels, leading us to conclude that the linear decision boundary was simply insufficient.

**CONCLUSIONS**

Radial Basis Kernels proved to be the most successful model, beating our baseline by an impressive 5%. When performing a Grid Search to select the top 3 most informative features via a Random Forest Classifier for a feature reduced model, loss of accuracy was observed. This, coupled with analysis of a feature correlation map, dissuaded fears of potentially dependent features undermining the mathematical basis of the support vector machines. By examining various slack variables for a linear kernel, and observing consistently poor performance, we feel confident in saying that a simplified linear boundary is not sufficient in explaining the complexity of human airport satisfaction.

**ACKNOWLEDGMENT**

Thank you to Professor Ahmad for a great course on Machine Learning!

**FIGURES**

Figure 1: Training Results Across 10-Fold Validation for Logistic Regression Model

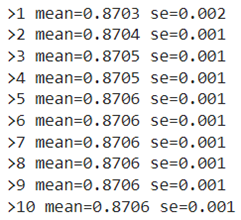
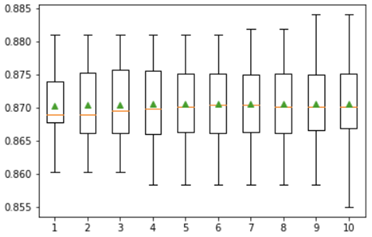


Figure 2: Correlation Heat Map of Features

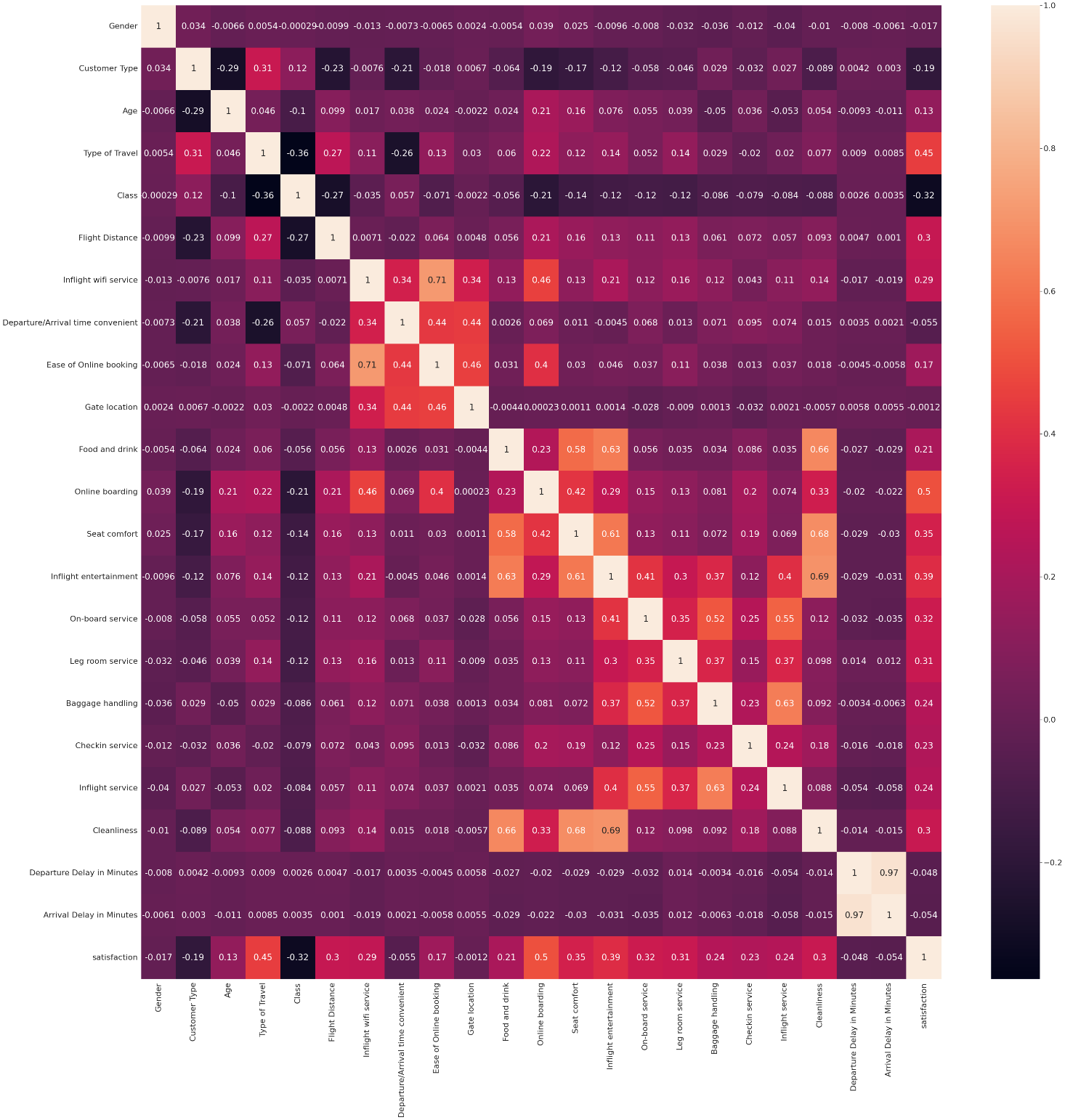


Figure 2



Figure 3

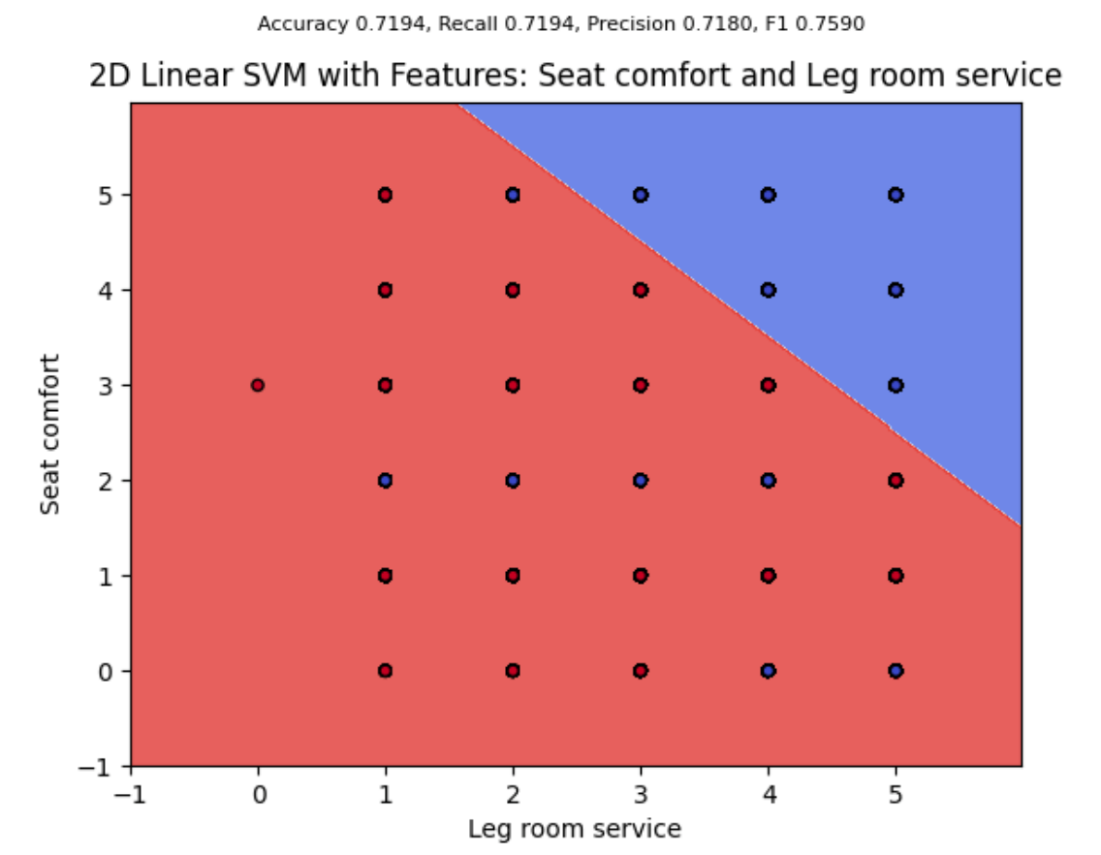


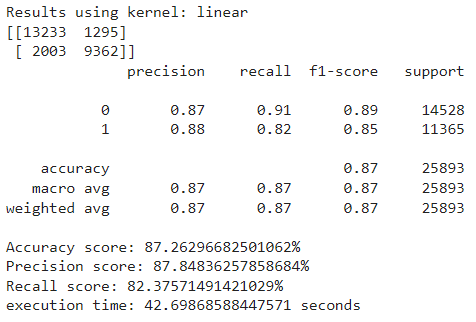
Table 1: Top 4 Performing Feature Combinations from Linear SVM (2 Feature Analysis)

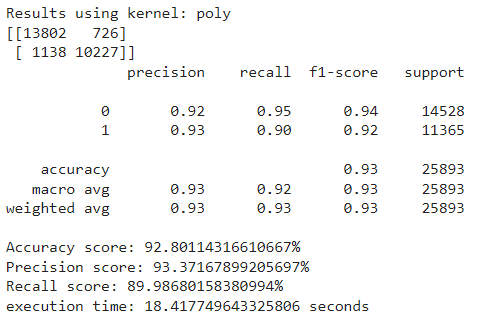
|  |  |  |
| --- | --- | --- |
| Features | Accuracy | F1 |
| Online Boarding and Inflight Service | 0.8804 | 0.8304 |
| Inflight Entertainment and Online boarding | 0.8109 | 0.8396 |
| Seat Comfort and Online Boarding | 0.7882 | 0.8006 |
| Inflight Entertainment and Leg Room Service | 0.7155 | 0.7457 |

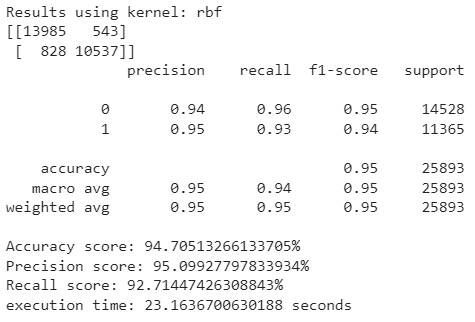
Table 2: Top 4 Performing Feature Combinations from Linear SVM (3 Feature Analysis)

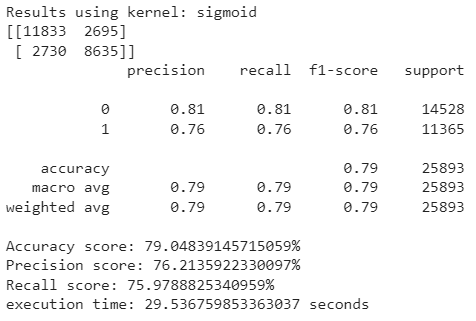
|  |  |  |
| --- | --- | --- |
| Features | Accuracy | F1 |
| Online Boarding, Inflight Entertainment and Inflight Wi-Fi | 0.8202 | 0.8434 |
| Online Boarding, Inflight Service and Inflight Wi-Fi | 0.8097 | 0.8338 |
| Online boarding, Leg Room and Inflight Service | 0.8064 | 0.8331 |
| Online boarding, Leg Room and Inflight Wi-Fi | 0.8025 | 0.8280 |

**ALL FEATURES**

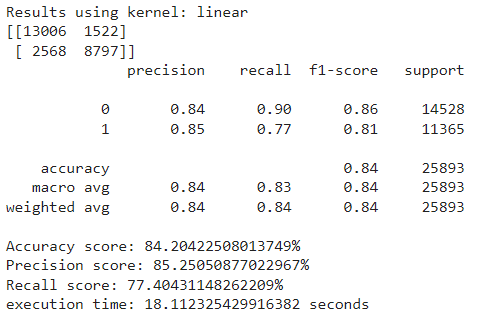


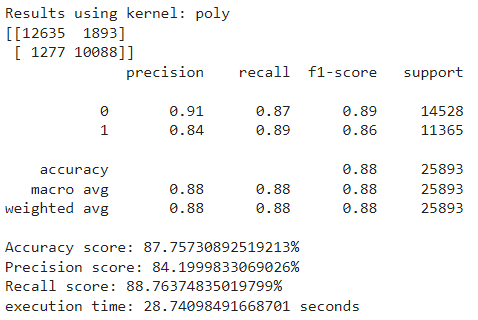


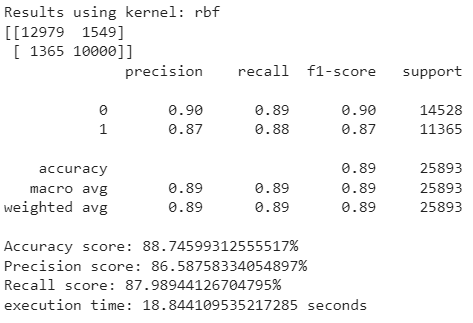


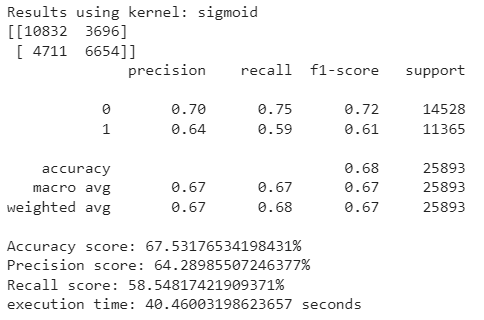


**TOP 3 FEATURES**









**REFERENCES**

1. Klein, TJ. **Airline Passenger Satisfaction**, Retrieved July 2022 from <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>.
2. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

Course Notes:

1. Ahmed, Bilal, "The Naïve Bayes Model", CS6140 Machine Learning Course Notes, Jan. 26, 2015.
2. Ahmed, Bilal, "Logistic Regression", CS6140 Machine Learning Course Notes
3. Ahmed, Bilal, "Support Vector Machines I", CS6140 Machine Learning Course Notes, Jan. 26, 2015.
4. Ahmed, Bilal, "Support Vector Machines III", CS6140 Machine Learning Course Notes, Jan. 26, 2015.
5. Pavlu, Virgil, "Kernels", CS6140 Machine Learning Course Notes, Nov. 30, 2014